



Research Article

<https://doi.org/10.1631/jzus.B2000494>



Integrative analysis of prognostic long non-coding RNAs with copy number variation in bladder cancer

Wenwen ZHONG¹, Dejuan WANG¹, Bing YAO¹, Xiaoxia CHEN², Zhongyang WANG¹, Hu QU¹, Bo MA¹, Lei YE¹, Jianguang QIU¹✉

¹Department of Urology, the Sixth Affiliated Hospital of Sun Yat-sen University, Guangzhou 510655, China

²Department of Medical Record Management Section, the Third Affiliated Hospital of Sun Yat-sen University, Guangzhou 510630, China

Abstract: Copy number variations (CNVs), which can affect the role of long non-coding RNAs (lncRNAs), are important genetic changes seen in some malignant tumors. We analyzed lncRNAs with CNV to explore the relationship between lncRNAs and prognosis in bladder cancer (BLCA). Messenger RNA (mRNA) expression levels, DNA methylation, and DNA copy number data of 408 BLCA patients were subjected to integrative bioinformatics analysis. Cluster analysis was performed to obtain different subtypes and differently expressed lncRNAs and coding genes. Weighted gene co-expression network analysis (WGCNA) was performed to identify the co-expression gene and lncRNA modules. CNV-associated lncRNA data and their influence on cancer prognosis were assessed with Kaplan-Meier survival curve. Multi-omics integration analysis revealed five prognostic lncRNAs with CNV, namely *NR2F1-AS1*, *LINC01138*, *THUMP3-AS1*, *LOC101928489*, and *TMEM147-AS1*, and a risk-score signature related to overall survival in BLCA was identified. Moreover, validated results in another independent Gene Expression Omnibus (GEO) dataset, GSE31684, were consistent with these results. Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis revealed that the mitogen-activated protein kinase (MAPK) signaling pathway, focal adhesion pathway, and Janus kinase-signal transducers and activators of transcription (JAK-STAT) signaling pathway were enriched in a high-risk score pattern, suggesting that imbalance in these pathways is closely related to tumor development. We revealed the prognosis-related lncRNAs by analyzing the expression profiles of lncRNAs and CNVs, which can be used as prognostic biomarkers for BLCA.

Key words: Bladder cancer; Copy number variation (CNV); Long non-coding RNA (lncRNA); Prognosis

1 Introduction

Of all the malignant tumors in humans, bladder cancer (BLCA) is the tenth leading cause of cancer-related morbidity. In 2018, nearly 540 000 people worldwide were diagnosed with BLCA and about 200 000 died from this disease (Bray et al., 2018). Although surgery, radiation therapy, chemotherapy, and individualized treatment significantly extend the overall survival including cancer-free survival of patients with BLCA, the five-year relative survival rate remains less than 50%, especially in metastatic patients (Hussain and James, 2003). A recent systematic review reported

a five-year recurrence rate for high-grade T1 BLCA of 42%, the progression rate of 21%, and a cancer-specific mortality rate of 13%, indicating that approximately two-thirds of the progression patients died from BLCA (Martin-Doyle et al., 2015). Recent molecular biology studies have pointed out that the occurrence and development of BLCA have heterogeneous molecular characteristics. The management of BLCA patients depends on the use of molecular detection techniques for diagnosis and for reliable decisions on potential targeted therapies (Mittra and Cote, 2009).

Copy number variation (CNV) is a mode of genetic structure variation that includes deletion, insertion, duplication, and complex multi-locus variation. CNV of DNA fragments ranges from kilobates (kb) to megabates (Mb) per million bases and is generally defined as the increase or decrease in the number of copies of a genome fragment between 1 and 3 Mb (Feuk et al.,

✉ Jianguang QIU, qiujiu@mail.sysu.edu.cn

Jianguang QIU, <https://orcid.org/0000-0002-3517-4110>

Received Aug. 30, 2020; Revision accepted Jan. 25, 2021;
Crosschecked June 15, 2021

© Zhejiang University Press 2021

2006; Redon et al., 2006; Nakamura, 2009). Changes in copy number in tumor genomes often lead to the expression of oncogenes and the inactivation of tumor suppressor genes, and have an important influence on the biological behavior of tumor cells (Henrichsen et al., 2009; Beroukhim et al., 2010). In a comprehensive study analyzing long non-coding RNA (lncRNA) in 12 cancer types with thousands of samples, researchers reported that more than 20% of lncRNA genes are located in regions with focal somatic copy number alterations (SCNAs) (Hu XW et al., 2014). By determining copy number, another study found that the oncogenic lncRNA *MALAT1* was overexpressed and that its up-regulation might be partially due to amplification of the *MALAT1* gene (Hu et al., 2015). A study by Hu Y et al. (2014) showed that lncRNA *GAPLINC* was associated with CNV in gastric cancer, promoting the invasiveness and poor prognosis of this cancer. CNV in the *IGHG3* gene at the 14q32.33 locus has been suggested to be associated with the high risk and high mortality of prostate cancer in African Americans (Ledet et al., 2013). CNV plays an important role not only in tumors, but also in other systemic diseases. In schizophrenia, for instance, about 25% of people with a chromosome 22q11.2 deletion, considered as a schizophrenia risk factor, have psychotic symptoms (Malhotra and Sebat, 2012). Another study suggested that the CNV region on chromosome 22q11.2 contains a lncRNA, *DGCR5*, which might be a potential regulator of genes associated with schizophrenia (Meng et al., 2018). Changes in copy duplication or deletion of the *HSP2A* gene were observed in infertile men with azoospermia (Eggers et al., 2015). The existence of a copy-number polymorphism in the upstream region of *IRGM* is associated with Crohn disease (McCarroll et al., 2008). Other studies have also reported the association of CNVs with diabetes, rheumatism, and inherited metabolic diseases (Potocki et al., 2007; The Wellcome Trust Case Control Consortium, 2010). Although CNVs are known to play critical roles in several diseases, including cancers, the regulatory relationship between CNV and lncRNA in BLCA has not been clarified.

lncRNAs are transcripts with length more than 200 nucleotides, with little or no protein-coding ability (Moran et al., 2012; Gudenat et al., 2019). An increasing number of studies have found differentially expressed lncRNAs in BLCA (Peter et al., 2014). Furthermore, lncRNAs play roles in inhibition or promotion of

tumor development and progression and are closely correlated with survival of and prognosis for BLCA (Nørskov et al., 2011; Kim et al., 2012; Martínez-Fernández et al., 2015). However, little research has been done to identify potential prognostic biomarkers by detecting DNA copy number amplifications and deletions in lncRNAs. In the present study, we were particularly interested in studying the relationship between whole-genome structural variations and lncRNAs in BLCA.

To investigate this relationship, we analyzed messenger RNA (mRNA) expression, DNA methylation, and DNA copy number data in detail and identified five molecular subtypes associated with prognosis in patients with BLCA. Differentially expressed mRNAs and lncRNAs in these five molecular subtypes were analyzed in BLCA tumors and compared to normal tissues. Furthermore, we studied the deregulation of lncRNAs due to copy number amplification or deletion in BLCA and performed Kaplan-Meier survival analysis of prognostic lncRNAs. Overall, our purpose was to determine the prognostic value of CNV-related lncRNAs in BLCA.

2 Materials and methods

2.1 Data source and processing

Methylation, RNAseq, CNV, and DNA mutation data and follow-up information on BLCA patients were downloaded from The Cancer Genome Atlas (TCGA) website (<https://portal.gdc.cancer.gov>). We first downloaded the fragments per kilobase of exon model per million mapped fragments (FPKM) and count data and then converted FPKM to transcripts per kilobase of exon model per million mapped reads (TPM) data. In this study, long intergenic non-coding RNA (lincRNA), sense-intronic, sense-overlapping, antisense, processed-transcript, and 3'-overlapping-ncRNA genes were classified as lncRNAs based on genecode file V22. Thereafter, we extracted the expression spectrum of lncRNAs and protein-coding genes (PCGs). Records of 423 BLCA patients containing 450 000 methylation data were screened for removal of probes whose sequences were not detected, cross-reactive CpG sites (Chen YA et al., 2013), and unstable CpG sites existing in the sex chromosomes and single nucleotide sites. NA probes mean the unspecific signal value that we are unable to determine as signal instability in the chip detection process. CNV data and single nucleotide mutation data were

processed for all samples; CNV data with germline differences were removed using MuTect software (Cibulskis K et al., 2013). After preparation of the data, we used the R package “icluster” to integrate PCG, methylation, and CNV data of 408 patients for cluster analysis with the cluster number set to 5 to identify stable sample subtypes (Shen et al., 2009).

2.2 Survival analysis and differentially expressed lncRNAs and genes in subtypes

To better classify the samples, we analyzed the influence of coding genes, CNVs, and methylation on prognosis for BLCA and established a model with univariate Cox proportional hazards regression analysis. The *P* threshold for significance was 0.05. We further screened differently expressed lncRNAs and coding genes between the different subtypes of tumor samples and normal samples using the R package “edgeR” (Robinson et al., 2010). During data processing, we first eliminated the genes with an average count of less than 1 in the expression profile and set fold change of >2 and false discovery rate (FDR) of 0.05 as the threshold. Furthermore, according to the differential multiple of lncRNAs in each subtype, we used the absolute value of the differential multiple as the rank of row to perform the gene set enrichment analysis (GSEA) (Subramanian et al., 2005).

2.3 Weighted gene co-expression network analysis (WGCNA)

We used the WGCNA algorithm (Langfelder and Horvath, 2008) to find the co-expression gene modules of the identified differentially expressed genes and lncRNAs. We screened samples as outliers whose distance was more than 20 000 and used Pearson’s correlation coefficient to calculate the distance between the gene and lncRNA. In the WGCNA, the soft threshold value was 4. Based on topological overlap matrix (TOM), we used average-linkage hierarchical clustering to cluster the genes. According to the criteria of hybrid dynamic shear tree, we set the minimum number of genes per lncRNA network module to 30. After determining the gene modules by the dynamic shear method, we calculated the eigenvectors for each module in turn, and then clustered the modules and merged the near modules into new modules cut on the basis of height=0.25, deepSplit=2, and minModuleSize=30. Functions of the modules enriched with lncRNAs were identified by

Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis.

2.4 Copy number expression profiles of lncRNAs

The copy number spectrum of lncRNAs was extracted from copy number data of 423 cases of BLCA downloaded from TCGA using GISTIC 2.0 software (Mermel et al., 2011). The copy number spectrum of lncRNAs was extracted by considering copy number >1 as the threshold for multiple copies and <-1 as the threshold for copy deletion. The ratio of multiple copies and copy deletion for each lncRNA was determined and the distribution of lncRNAs in the genome was observed. Furthermore, we calculated the correlation distribution between the lncRNA expression profile and copy number and identified frequently changing regions in the genome of BLCA patients using the logistic algorithm. To observe the relationship between lncRNA expression and copy number, we chose lncRNA with over 15% copy number ratio in each sample. For further analysis, each lncRNA in the copy amplification or copy missing and copy normal samples was selected, for which the expression level was more than 0 in at least ten samples in each group.

2.5 Prognostic biomarkers of copy number variation-related lncRNAs in bladder cancer

To systematically identify lncRNAs that can be used as prognostic markers, we analyzed the copy number of different lncRNAs in each subtype and selected those lncRNAs that differed in each subtype. Further, we analyzed the relationship between lncRNAs and overall survival by univariate Cox analysis. The selection *P* value was 0.01. Moreover, we calculated the risk score for each sample according to the expression level of the sample and plotted a receiver operating characteristic curve (ROC) to analyze the predictive classification efficiencies of 1-, 3-, and 5-year progressions. To verify the prognostic significance of these CNV-related lncRNAs, we downloaded the GSE31684 data from GPL570.

3 Results

3.1 Identification of prognosis-related molecular subtypes

We screened 408 patients after comprehensive analysis of the data for coding genes, CNV, and

methylation with respect to BLCA prognosis based on a univariate Cox proportional hazards model. A total of 3214 coding genes, 6745 CNV regions, and 20207 CpG loci were found. Subsequently, we identified five subtypes cluster 1 (C1), C2, C3, C4, and C5 consisting of 54, 60, 45, 190, and 59 samples, respectively. Further, we analyzed the prognostic differences among these five subtypes with regard to BLCA. These five subtypes showed significant prognostic differences, with a log-rank *P* value of 0.0003; the C1 group had the best prognosis (Fig. 1a). The details of gene mutations in BLCA were also analyzed, and the top 20 genes with the highest proportion of mutations were screened and visualized (Fig. 1b). It was observed that mutations in *TP53*, *TTN*, *KMT2D*, *MUC16*, and *KDM6A* were more common than those in the other genes, indicating that the high mutation frequency in these genes plays an important role in the development of cancer. Overall, these findings suggest that combined analyses of genomic and transcriptome data can reveal their regulatory relationship and help in predicting different outcomes in patients with BLCA.

3.2 Screening of differentially expressed lncRNAs and genes in subtypes

Next, we determined the differentially expressed PCGs and lncRNAs in tumors and adjacent tissues and in the five subtypes (C1, C2, C3, C4, and C5) (Table 1). A total of 2507 different lncRNAs and 3453 PCGs were obtained from the five subtypes, with 911, 711,

957, 636, and 701 lncRNAs in the C1, C2, C3, C4, and C5 subtypes, respectively. The differences in lncRNAs among the subtypes are shown in Figs. 2a–2e and the numbers of lncRNA and PCGs are presented in Fig. 2f. The C1 and C3 samples were observed to have the maximum differences in lncRNAs and the number of PCGs, whereas the C4 samples had the fewest differences. Also, the number of up-regulated lncRNAs in each subtype was lower than that of down-regulated lncRNAs.

In addition, we downloaded data for 717 disease-related lncRNAs from the LncRNADisease (Chen G et al., 2013) and Lnc2Cancer databases (Ning et al., 2016) and compared them with the differentially expressed lncRNAs in the five subtypes (Fig. 2g). Among these data, 128 lncRNAs were closely correlated with various diseases. A hypergeometric test was used at a *P* threshold of 0.0001. We also performed GSEA, and the results revealed that the differential lncRNAs were concentrated in the gene set with a higher differential multiple (Figs. 3a–3e). The intersection of lncRNAs between the five subtypes as shown in Fig. 3f, showed that several lncRNAs were common in the five subtypes.

3.3 Co-expressed modules between lncRNAs and protein-coding genes based on WGCNA

We used the WGCNA co-expression algorithm to mine co-expressed gene and lncRNA modules based on the different PCGs and lncRNA expression spectrum.

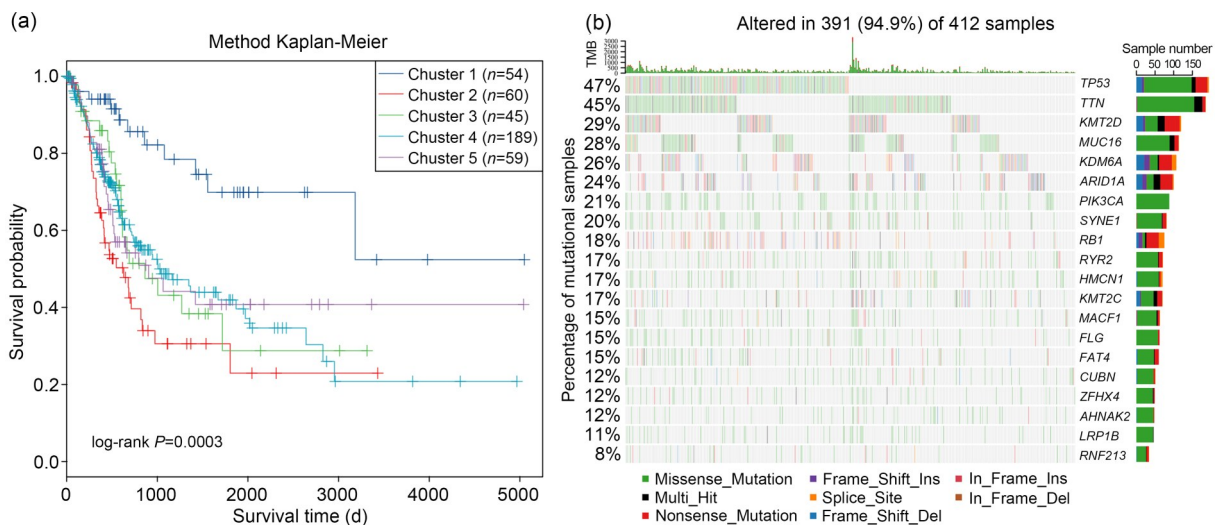


Fig. 1 Prognosis of five subtypes and analysis of mutant genes. (a) Kaplan-Meier curve for five subtypes; (b) Mutation landscape for the top 20 genes in bladder cancer. TMB: tumor mutational burden.

Table 1 Differentially expressed protein-coding genes and lncRNAs between tumors and adjacent tissues (subtype All) and five subtypes (C1, C2, C3, C4, and C5)

Type	C1	C2	C3	C4	C5	All
PCG_Down	1671	1438	1810	1271	1442	2285
PCG_Up	1023	796	1173	678	790	1492
PCG_All	2694	2234	2983	1949	2232	3777
Lnc_Down	605	512	632	450	498	901
Lnc_Up	306	199	325	186	203	475
Lnc_All	911	711	957	636	701	1376

PCG: protein-coding gene; Lnc: long non-coding RNA (lncRNA); C1–C5: cluster 1–cluster 5.

Expression profiles of lncRNAs and PCGs were extracted and the samples were analyzed by hierarchical clustering (Fig. 4a). Finally, 412 samples were obtained. To ensure that the web was a scale-free network, we chose soft-thresholding power β as 4 (Figs. 4b and 4c). Based on the statistics mentioned in Section 2.3, we obtained 25 modules (Fig. 4d). It should be noted that the grey module is a collection of genes that cannot be aggregated into other modules. The details of PCGs and lncRNAs in each module are summarized in Table 2. Fig. 4e shows a significant enrichment of lncRNAs in brown, yellow, and pink modules, where the P value indicates significant aggregation of lncRNAs in the

module and fold change (FC) indicates the aggregation multiple of lncRNAs. In addition, three modules (brown, yellow, and pink) were selected for functional analysis by using the R software package “clusterProfiler” (Yu et al., 2012) with a P threshold value of 0.05. As a result, these three modules were enriched in 41 KEGG pathways and the modules tended to be enriched in different pathways, suggesting different functions for each module (Fig. 5a). Twelve KEGG pathways, including cyclic adenosine monophosphate (cAMP) and glucagon signaling pathways, were enriched in the brown module (Fig. 5b). Ten pathways were enriched in the pink module (Fig. 5c), whereas the 19 pathways closely related to tumorigenesis, including VEGF and p53 signaling pathways, were enriched in the yellow module.

3.4 Identification of copy number variation-related lncRNAs in bladder cancer

First, we obtained the spectrum of lncRNAs with CNV data and described the ratios of multiple copies and copy deletions of lncRNAs along with their distribution on the genome (Fig. 6a). Second, assessment of the distribution of Pearson’s correlation between the lncRNA expression and CNV suggested a positive correlation

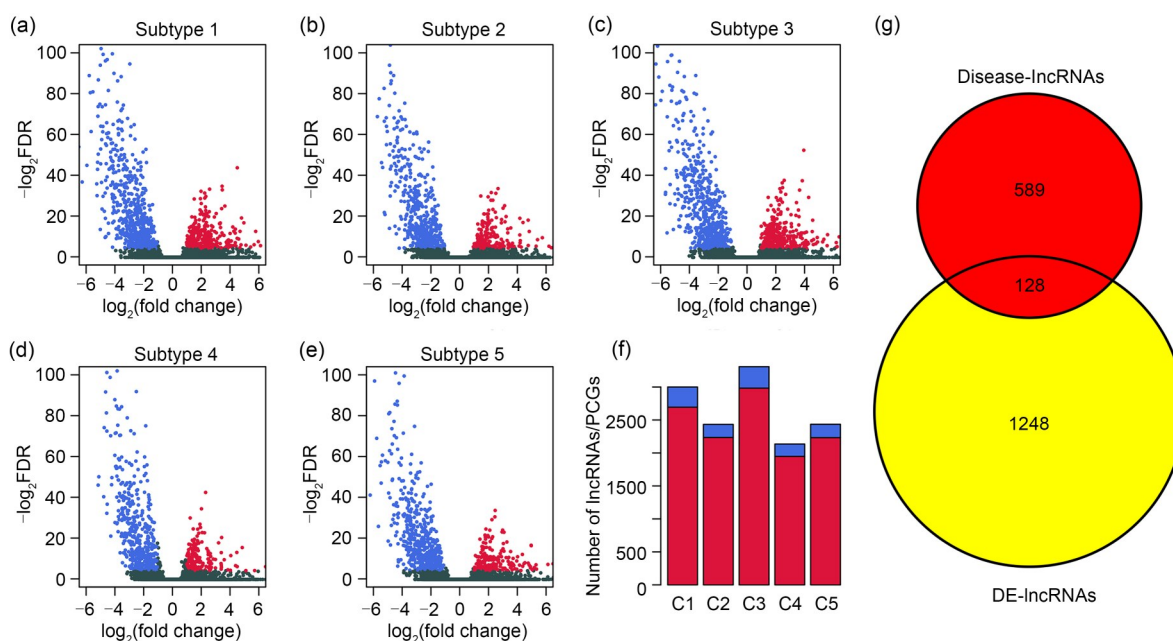


Fig. 2 Distribution of differentially expressed long non-coding RNAs (DE-lncRNAs) and protein-coding genes (PCGs) among five subtypes. (a–e) Volcanic maps of DE-lncRNAs in the five subtypes (red denotes up-regulated lncRNAs, and blue represents down-regulated lncRNAs); (f) Composition of DE-lncRNAs and PCGs in five subtypes (blue represents DE-lncRNAs and red represents PCGs); (g) Venn diagram of the intersection of DE-lncRNAs (yellow) and disease-lncRNAs (red). FDR: false discovery rate.

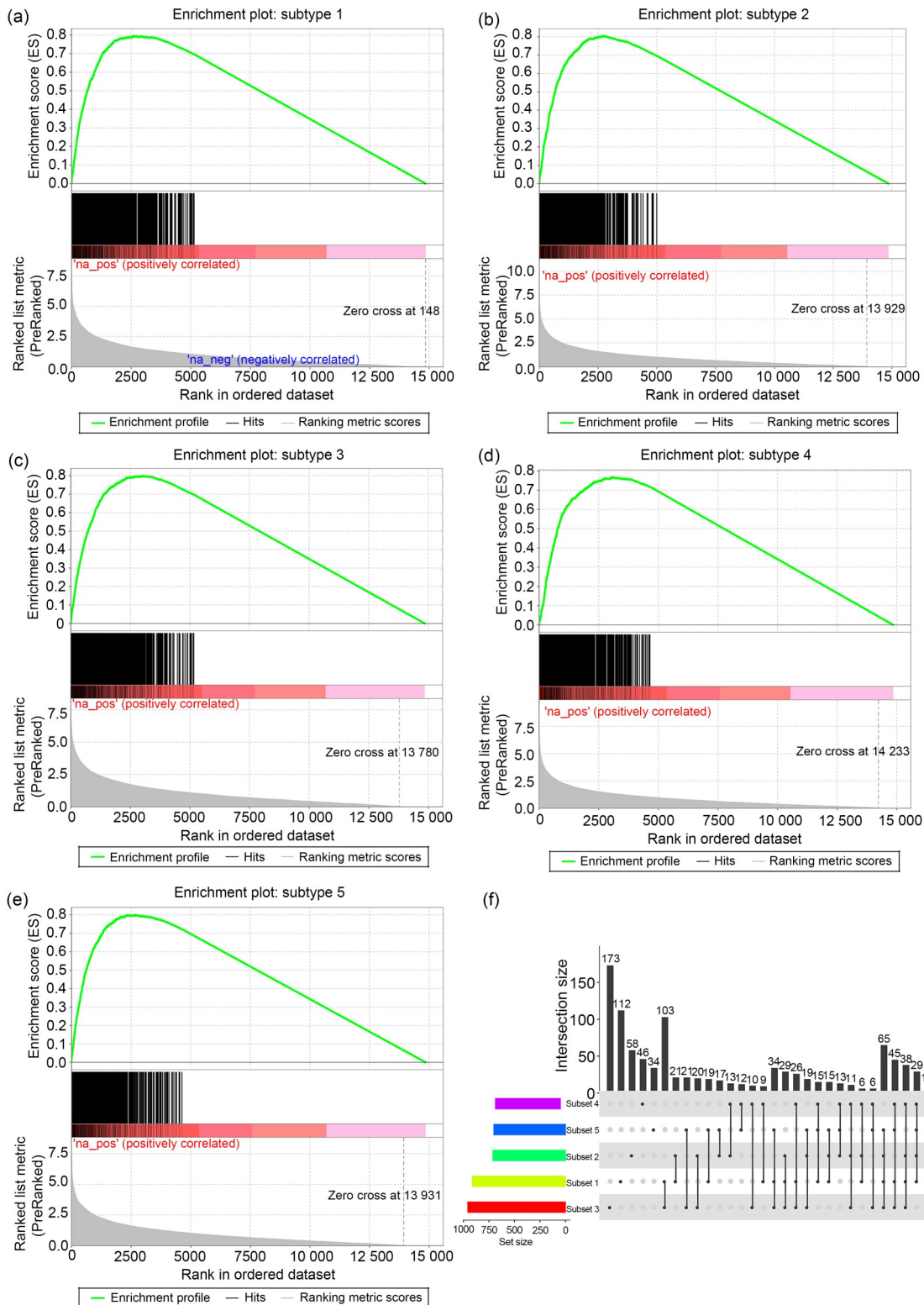


Fig. 3 Distribution states of differently expressed long non-coding RNAs (DE-lncRNAs) between subtypes. (a–e) The results of gene set enrichment analysis (GSEA) in each subtype according to the rank of difference multiples. GSEA shows different enrichment states among the five identified subtypes (subtypes 1, 2, 3, 4, and 5). (f) The intersection of DE-lncRNAs of the five subtypes. Dots indicate subtype and lines represent overlapped lncRNAs in subtypes. The lncRNA size points towards the amount of DE-lncRNAs.

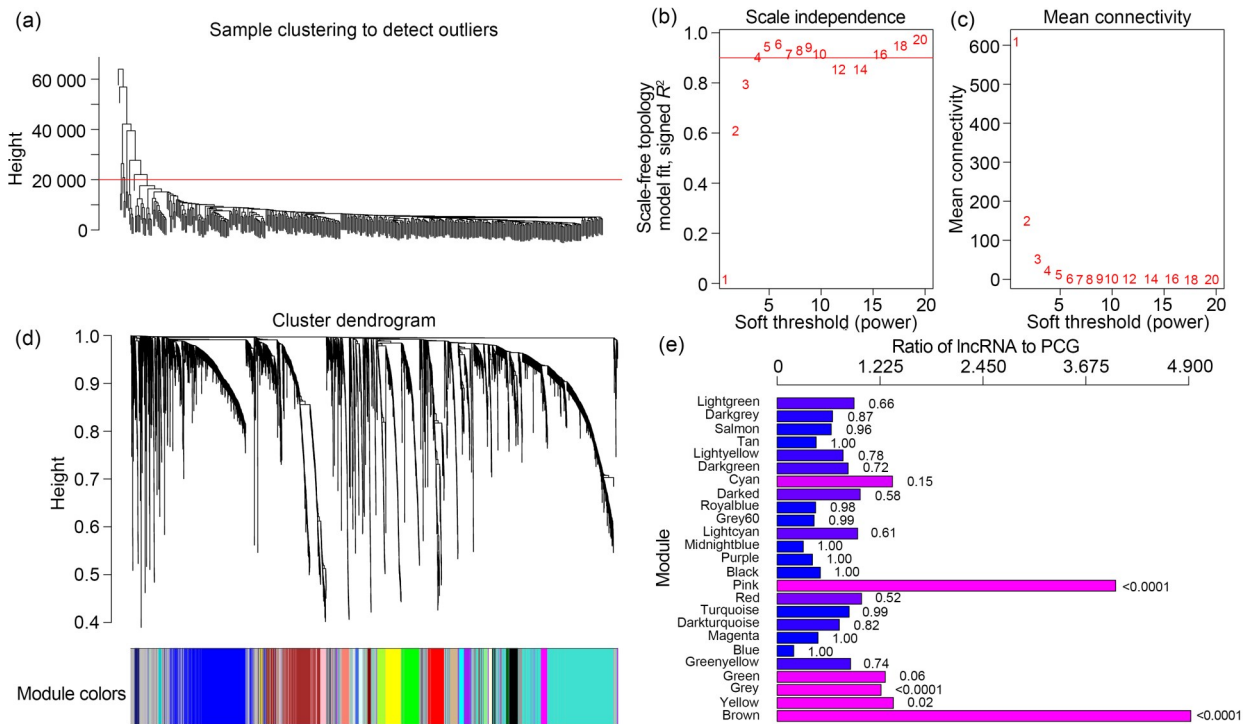


Fig. 4 Identification of co-expressed modules using weighted gene co-expression network analysis (WGCNA). (a) Cluster analysis is performed to detect outliers. Objects with height greater than 20 000 (above the red line) are excluded. (b, c) Analysis of network topology for various soft-thresholding powers. The red line represents the square of correlation coefficient at 0.9. (d) Cluster dendrogram of samples and modules is presented with different colors. (e) Bar chart of relative multiples of long non-coding RNA (lncRNA)/protein-coding gene (PCG) ratio among 25 modules. The value on the right side represents the P value.

between these factors; the distribution was significantly higher than random distribution, with a P value of 2.2×10^{-16} (Fig. 6b). Further, we analyzed the copy number of lncRNAs. Amplification of lncRNAs was significantly greater than deletions (Figs. 6c and 6d), suggesting that copy variation in lncRNAs might be associated with the development of BLCA. To further assess the relationship between lncRNA expression and CNVs, we selected eight lncRNAs with a CNV ratio of >15% in each sample (Fig. 7) and observed that amplification was significantly higher in five of these lncRNAs (*LOC729867*, *LOC101928372*, *CASC15*, *FLJ42969*, and *UBR5-AS1*) than the normal copy number of other lncRNAs in the sample. These results suggest that lncRNA expression is closely related to CNV.

To systematically identify the prognostic markers for CNV-related lncRNAs, we selected lncRNAs that differed in the five subtypes and analyzed their correlation with overall survival in BLCA by univariate Cox regression analysis with a P selection threshold value of 0.01. We screened five lncRNAs found to be correlated with overall survival (Fig. 8) among 23

lncRNAs associated with significant prognosis. The details of expression and CNV of these five lncRNAs, including *LINC01138*, *THUMP3-AS1*, *NR2F1-AS1*, *LOC101928489*, and *TMEM147-AS1*, at different stages of BLCA are shown in Fig. S1. As seen in Fig. S1, the expression of *NR2F1-AS1* and *LOC101928489* differed at different BLCA stages, with significant differences in CNVs of *LINC01138* and *THUMP3-AS1* between stages III and IV, and in CNV of *NR2F1-AS1* between stages II and IV. Multivariate Cox analysis of these five lncRNAs is presented in Table 3. We can conclude the risk-score model as follows: $\text{RiskScore} = -(0.10362 \times \exp^{\text{LINC01138}}) - (0.01243 \times \exp^{\text{THUMP3-AS1}}) + (0.11439 \times \exp^{\text{NR2F1-AS1}}) + (0.93572 \times \exp^{\text{LOC101928489}}) - (0.01938 \times \exp^{\text{TMEM147-AS1}})$, where \exp is expression (e.g., $\exp^{\text{LINC01138}}$ is the expression of *LINC01138*).

3.5 Analysis of area under the curve of the risk-score model

We calculated the risk score for each sample based on expression level and plotted the sample risk-score

Table 2 Protein-coding genes (PCGs) and long non-coding RNA (lncRNA) expression in each module

Module	All	Lnc	PCG	P value	FC
Brown	475	305	170	4.3×10^{-73}	4.925
Yellow	200	67	133	0.018	1.383
Grey	979	304	675	0.000	1.236
Green	188	60	128	0.061	1.287
Greenyellow	87	21	66	0.744	0.873
Blue	807	54	753	1.000	0.197
Magenta	113	17	96	0.999	0.486
Darkturquoise	33	7	26	0.818	0.739
Turquoise	1102	262	840	0.994	0.856
Red	168	45	123	0.520	1.004
Pink	116	69	47	0.000	4.030
Black	165	26	139	1.000	0.513
Purple	98	13	85	1.000	0.420
Midnightblue	59	6	53	1.000	0.311
Lightcyan	58	15	43	0.608	0.958
Grey60	58	8	50	0.994	0.439
Royalblue	35	5	30	0.975	0.457
Darkred	34	9	25	0.578	0.988
Cyan	63	21	42	0.146	1.372
Darkgreen	34	8	26	0.724	0.845
Lightyellow	36	8	28	0.785	0.784
Tan	83	12	71	0.998	0.464
Salmon	79	15	64	0.959	0.643
Darkgrey	31	6	25	0.874	0.659
Lightgreen	52	13	39	0.661	0.915

P value indicates a significant aggregation of lncRNA in the module; fold change (FC) indicates the aggregation multiple of lncRNA. Lnc: lncRNA.

distribution as shown in Fig. 9a. It can be seen from the graph that the overall survival of the samples with a high-risk score was significantly lower than that of the samples with a low-risk score, suggesting that patients with a high-risk score have poorer survival. Levels of expression of the five different prognosis-related lncRNAs varied with increase in risk score. High expression of *NR2F1-AS1* and *LOC101928489* was associated with high risk, indicating that these two genes are risk factors, whereas high expression of *LINC01138*, *THUMPD3-AS1*, and *TMEM147-AS1* was correlated with low risk, indicating that these genes act as protective factors. Using the R software package “timeROC” (Blanche et al., 2013), we analyzed the ROCs of 1-, 3-, and 5-year survivals with the risk-score model resulting in

a high area under the curve (AUC) of >0.64 (Fig. 9b). Finally, we converted the risk score to z-score and divided the samples into high-risk and low-risk groups when the z-score was higher or lower than zero, respectively. As a result, 196 samples were classified as high risk and 209 were classified as low risk. The survival curve shown in Fig. 9c suggests that patients in the high-risk group had a significantly worse prognosis, with a P value of <0.0001 and a hazard ratio (HR) of 2.312.

To confirm the constructed risk-score model using the five CNV-related lncRNAs relevant to BLCA prognosis, we downloaded GSE31684 data from the GPL570 platform as a test dataset for analysis. We compared these five lncRNAs to the probe and found that annotations for only four lncRNAs could match the probe. Using the same model and the same coefficients in the training set, we calculated the risk score of each sample according to the sample expression level and plotted the risk-score distribution (Fig. 10a), based on which we can draw similar conclusions as obtained using the training set. High expression of *NR2F1-AS1* and *THUMPD3-AS1* could be identified as risk and protective factors, respectively, as in the training set. Similarly, we analyzed the 1-, 3-, and 5-year survival efficiencies (Fig. 10b). Finally, we divided the samples into high-risk and low-risk groups using the same methods; 28 samples were classified as high risk and 65 samples were classified as low risk. The Kaplan-Meier curve demonstrates that patients in the low-risk group had a better survival prognosis than those in the high-risk group, with an HR of 1.981 (Fig. 10c).

3.6 Potential regulatory pathways of the risk-score model

We used the R software package “GSVA” (Hänzelmann et al., 2013) to perform a single-sample GSEA (ssGSEA) to find the gene ontology (GO) enrichment function of the five distinctive lncRNAs. This analysis (Fig. S2) indicated similar biological function, cellular component, and molecular function. To observe the relationship between biological function and the calculated risk scores of the five lncRNAs in different samples, we selected the corresponding gene expression profiling samples to perform ssGSEA. ssGSEA scores for each function were obtained by calculating the scores of different functions for each sample. The correlations between these functions and risk score were calculated, yielding a correlation coefficient greater than 0.4

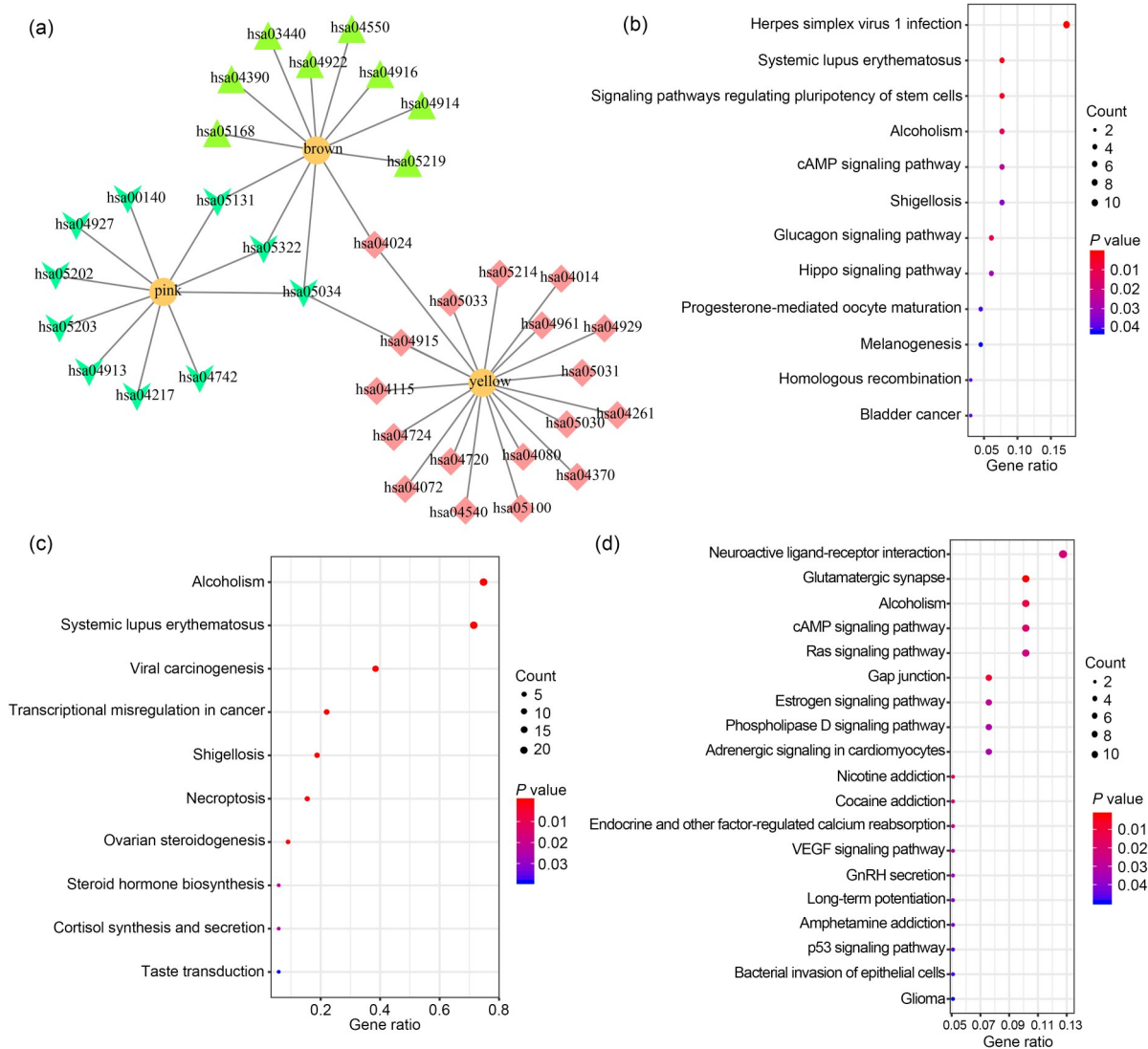


Fig. 5 Correlation analysis between modules. (a) Network relationship between the enrichment results of the three modules; (b) Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment result of brown module; (c) Gene Ontology (GO) enrichment analysis of pink module; (d) GO enrichment analysis of yellow module. cAMP: cyclic adenosine monophosphate; VEGF: vascular endothelial growth factor; GnRH: gonadotropin-releasing hormone.

(Fig. 11a). Most functions were positively correlated with risk score, although a few showed a negative correlation. Seventeen KEGG pathways were selected with a correlation coefficient greater than 0.4. Cluster analysis was performed according to the enrichment score for these pathways (Fig. 11b). Among these 17 pathways, scores of the mitogen-activated protein kinase (MAPK) signaling pathway, Janus kinase (JAK) signaling pathway, and others increased along with risk score, suggesting that imbalance in these pathways is closely related to tumor development in BLCA.

4 Discussion

Despite the rapid development in cancer medicine in recent years, the prognosis for BLCA patients, especially those in advanced cancer stages, has not been improved. Because of the high heterogeneity of tumors, therapeutic outcomes are not satisfactory. Therefore, development of molecular markers to predict tumor prognosis is of great clinical importance. In recent years, scientists have used second-generation sequencing technology to fully understand the complex biological

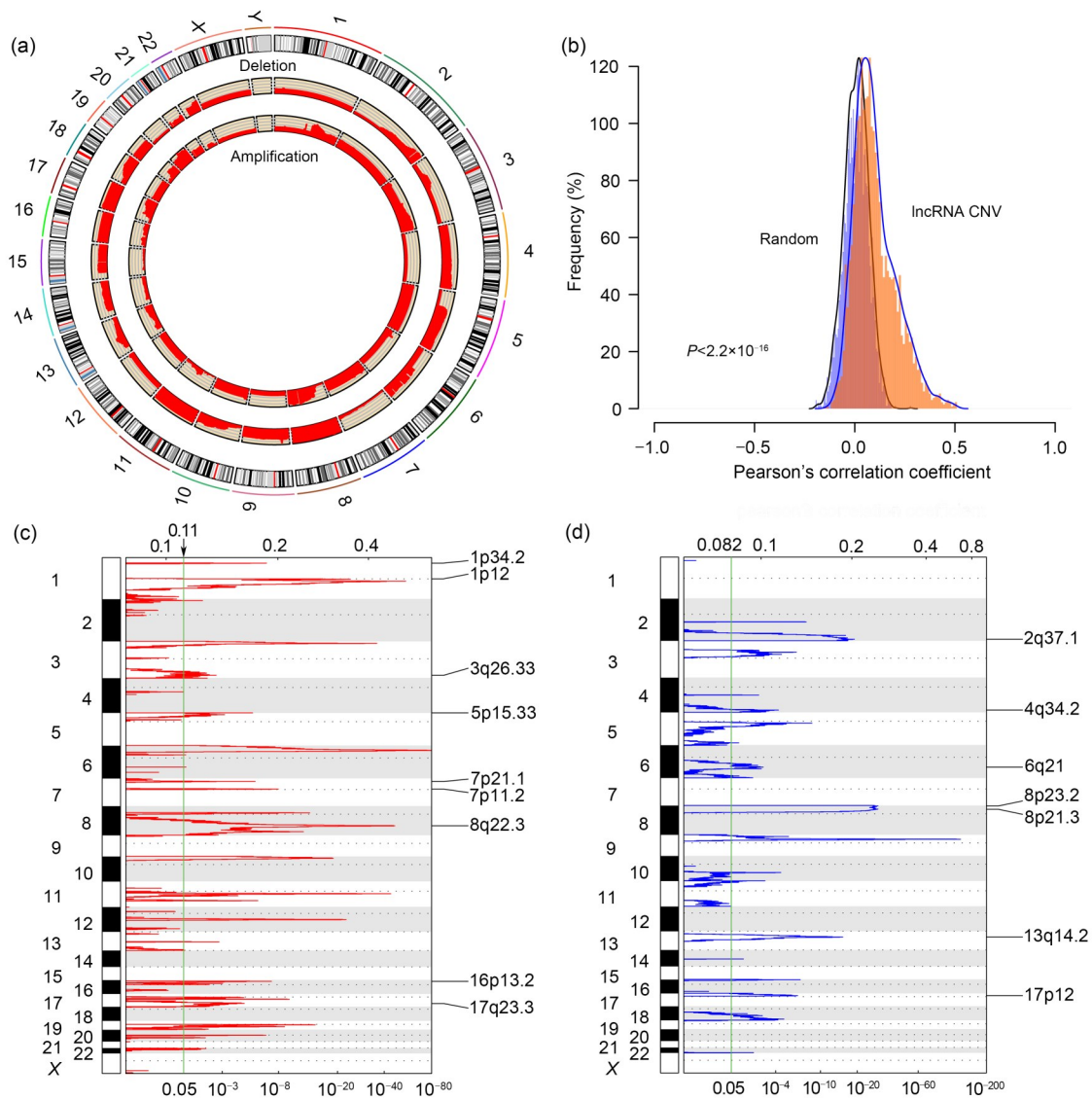


Fig. 6 Pattern of copy number variation (CNV) profiles in the whole genome. (a) Proportional distribution of copy deletion and copy amplification of lncRNAs in the genome; (b) The relative distribution of long non-coding RNA (lncRNA) expression and CNVs. Light blue represents random distribution and orange represents the actual distribution. The *t*-test was used to test the difference between the two distributions. (c) The regions where lncRNA copy number is amplified in the genome. (d) The regions where lncRNA copy number is deleted in the genome. (c, d) The Y-axis stands for genome position; X-axis stands for false-discovery rates (*q* values) (bottom) and scores from GISTIC 2.0 (top).

behaviors of tumors. Cluster analysis has also been widely used to discern tumor heterogeneity. Cluster analysis is a process of classifying data into different clusters, with objects in the same cluster having greater similarity. Statistically, cluster analysis is a good method to simplify data through data modeling. Using cluster analysis of RNA-seq, DNA mutation, methylation, and CNV data present in TCGA database, this study found five molecular subtypes (C1, C2, C3, C4, and C5) correlated with BLCA prognosis. Of these subtypes, the C1

subtype was correlated with the best prognosis. Each of these five subtypes had different molecular characteristics; nonetheless, there were several commonalities among these subtypes. In addition, module analysis of differential lncRNAs revealed that the lncRNAs were mainly clustered in three different functional modules. Further analysis of expression of lncRNAs with respect to CNV showed that expansion or deletion of RNA copy number can affect the expression level. Finally, we identified five CNV-related lncRNAs that were closely

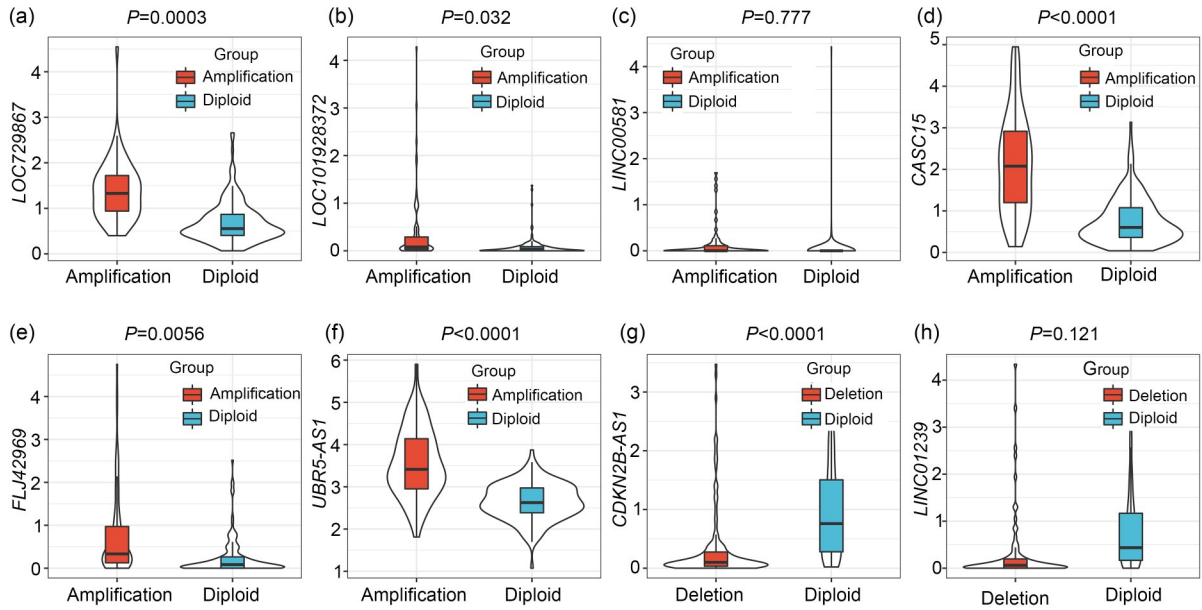


Fig. 7 Eight long non-coding RNAs (lncRNAs) with a copy ratio greater than 15% in each sample induced by copy number amplification or deletion. Blue (diploid) represents normal copy, and red (deletion or amplification) represents variant copy.

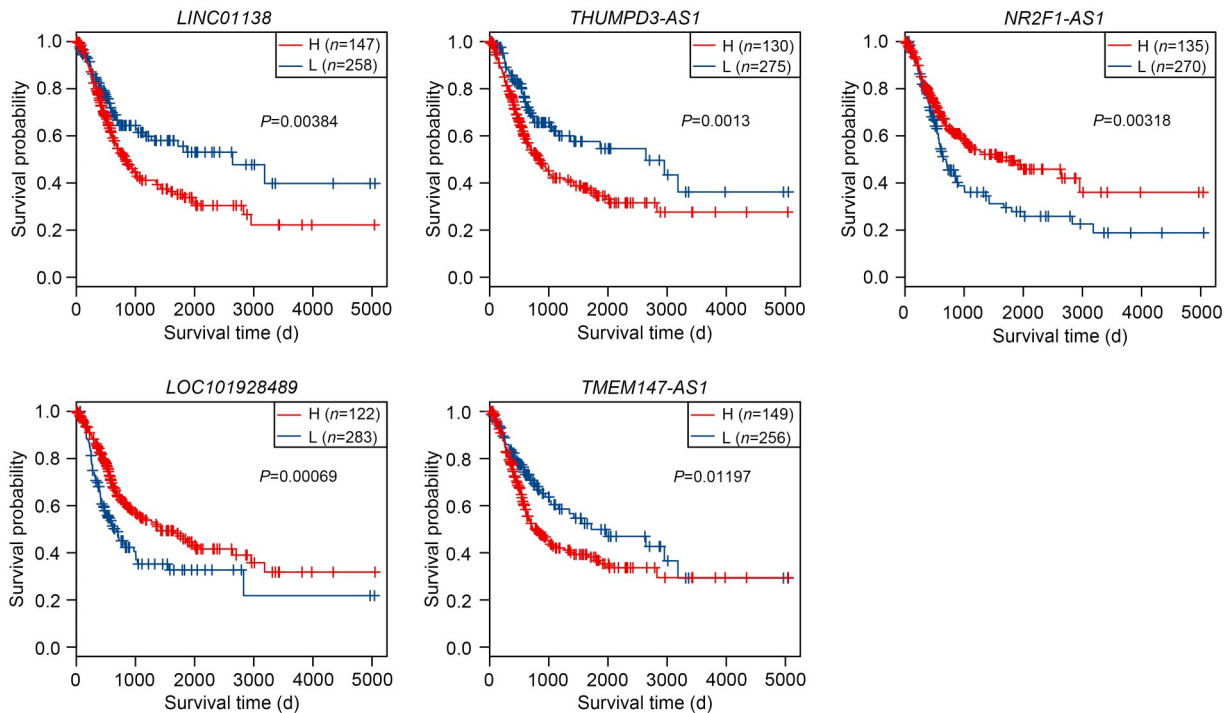


Fig. 8 Kaplan-Meier survival curve of five prognostic copy number variation (CNV)-related long non-coding RNAs (lncRNAs). H: high expression; L: low expression.

associated with survival of BLCA patients; a risk-score model composed of these lncRNAs successfully predicted BLCA prognosis and was also useful for validation using another Gene Expression Omnibus (GEO) dataset.

Carcinogenesis is a complex process that usually involves thousands of genomic alternations, including single nucleotide mutations and CNVs. CNVs can cause a series of functional changes, such as gene-dose effects, gene disruption, gene fusion, and location-change effects.

Table 3 Five long non-coding RNAs (lncRNAs) with significant prognosis among five subtypes

Gene	Coefficient	HR	z-score	P value	95% CI
<i>LINC01138</i>	-0.103 62	0.9016	-2.490	0.0128	0.831–0.978
<i>THUMPD3-AS1</i>	-0.012 43	0.9876	-0.780	0.4353	0.957–1.019
<i>NR2F1-AS1</i>	0.114 39	1.1212	3.331	0.0009	1.048–1.199
<i>LOC101928489</i>	0.935 72	2.5490	4.294	1.76×10^{-5}	1.663–3.907
<i>TMEM147-AS1</i>	-0.019 38	0.9808	-1.240	0.2151	0.951–1.011

HR: hazard ratio; CI: confidence interval.

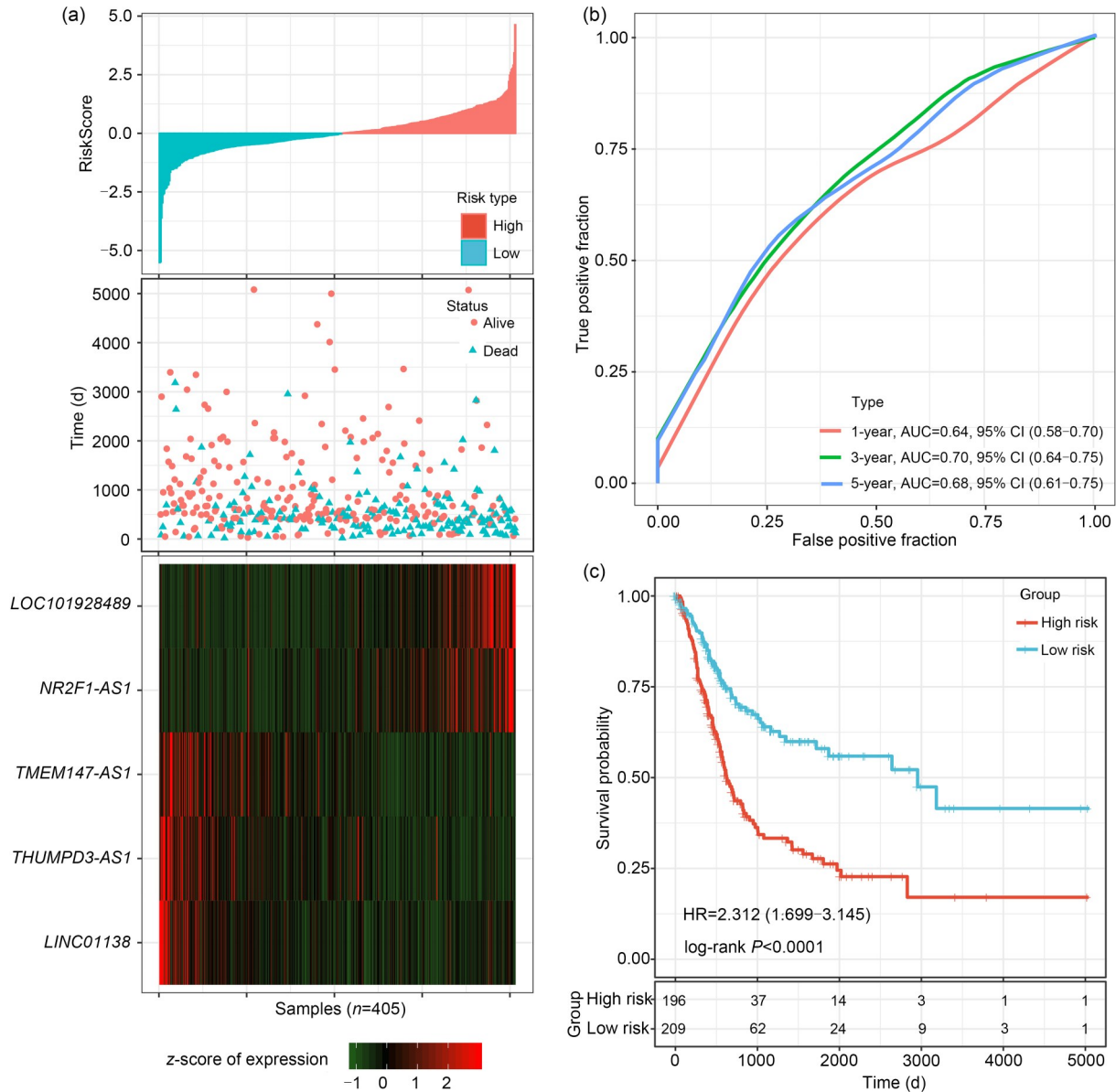


Fig. 9 Analysis of area under the curve (AUC) of the risk-score model in the training dataset. (a) The risk score for each sample based on expression level and sample risk-score distribution (top panel); patient survival status and duration based on the risk-score model constructed by the five copy number variation (CNV)-related long non-coding RNAs (lncRNAs) (median panel); heatmap of the five lncRNAs expression profiles in the training set (bottom panel). (b) AUCs of 1-, 3-, and 5-year survivals of BLCA patients predicted by the risk-score model. (c) Kaplan-Meier analysis of patients' overall survival in the high-risk and low-risk subgroups based on the CNV-related lncRNA risk-score model. HR: hazard ratio; CI: confidence interval.

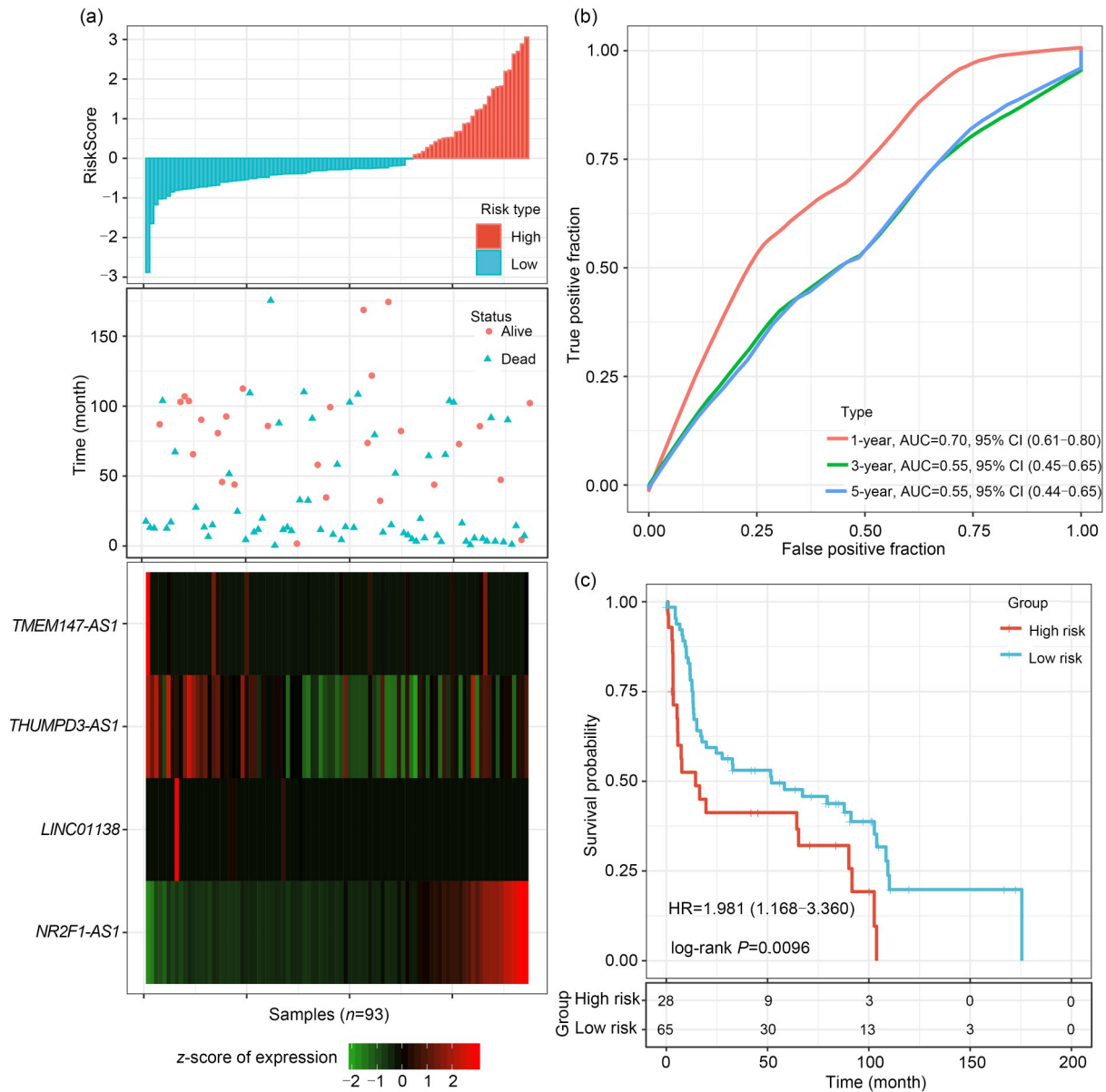


Fig. 10 Area under the curve (AUC) of the risk-score model in the validating set GSE31684. (a) The risk-score for each sample based on the expression level and the sample risk-score distribution (top panel); patient survival status and duration based on the risk-score model constructed by the copy number variation (CNV)-related long non-coding RNAs (lncRNAs) (median panel); heatmap of the four lncRNA expression profiles matched in GSE31684 (bottom panel). (b) AUCs of 1-, 3-, and 5-year survivals of BLCA patients. (c) Kaplan-Meier analysis of patients' overall survival in high-risk and low-risk subgroups in the validating set. HR: hazard ratio; CI: confidence interval.

Extensive changes in CNV and accompanying imbalances in gene expression may disrupt cell metabolism and can be decisive in determining physiological balance. For these reasons, CNV is closely related to the development of the disease. Studies have shown that CNVs can cause biological changes in various tumors, including esophageal carcinoma, gastric cancer, adrenocortical carcinoma, liver cancer, and BLCA. A clinical study

showed that platinum-resistant epithelial ovarian cancer patients with *TOP2A* gene copy number gain, which was paralleled by increased expression of the protein, showed better clinical response to pegylated-liposomal doxorubicin (Erriquez et al., 2015). CNV and overexpression of *PTP4A3* were found in esophageal carcinoma tissue samples by whole genome and DNA CNV analysis (Liu et al., 2018). Exome sequencing

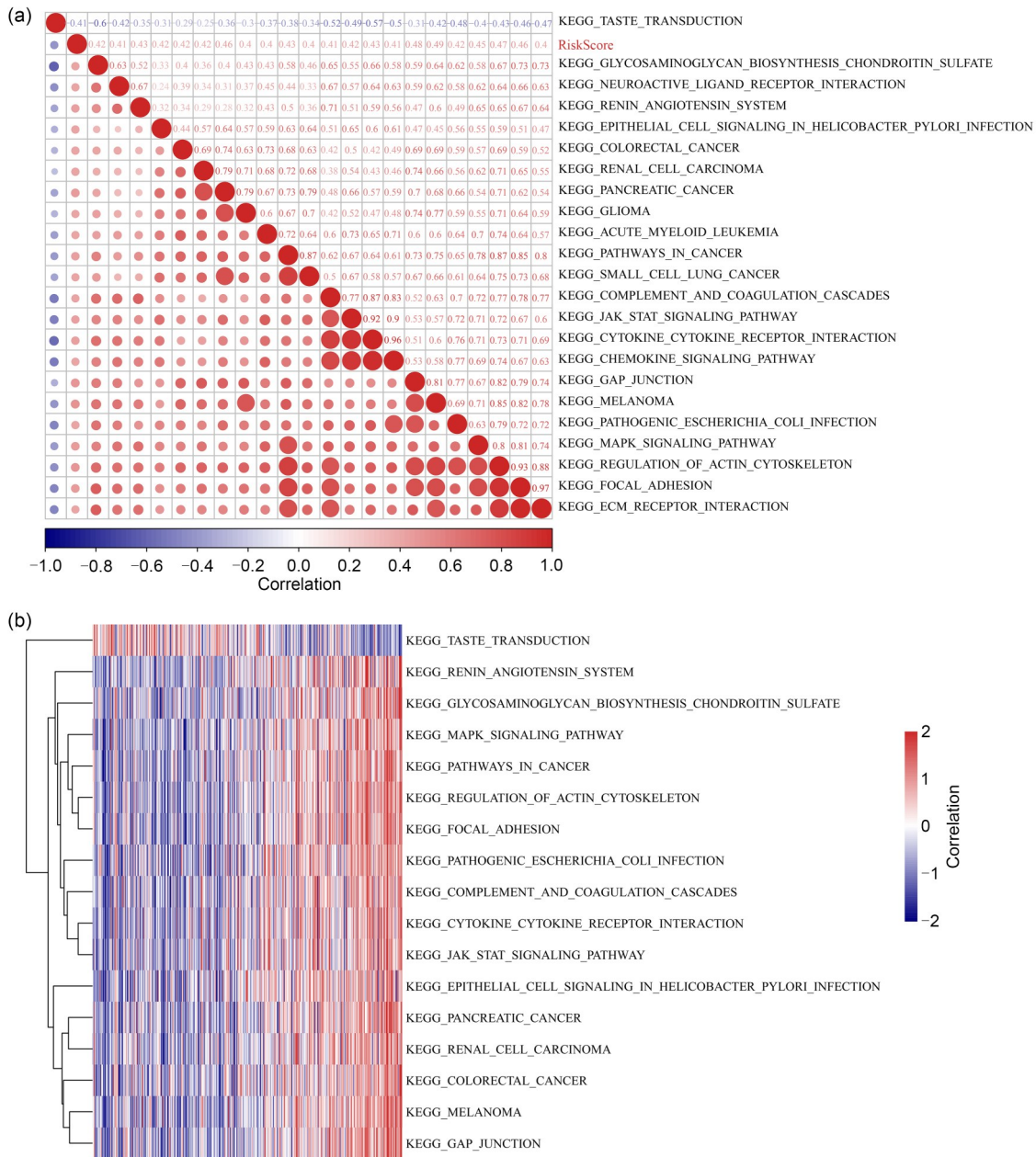


Fig. 11 Potential regulatory pathways of the risk-score model. (a) Clustering of correlation coefficients between Kyoto Encyclopedia of Genes and Genomes (KEGG) pathways and RiskScore with a risk score correlation greater than 0.4; (b) Single sample gene set enrichment analysis (ssGSEA) of KEGG pathways with a correlation of more than 0.4 when the risk score is varied. In the horizontal axis, the risk score increases from left to right.

revealed that more than half of adrenocortical carcinoma patients showed large-scale amplification of chromosome 19, which was associated with advanced disease (Rubinstein et al., 2016). CNVs have also been found in other cancers, including hepatocellular carcinoma (Xu et al., 2015; Chappell et al., 2016), pediatric sarcoma (Cheng et al., 2019), and squamous lung cancer (Chen

et al., 2019). A comprehensive genomic map in 295 clinically advanced urothelial BLCA patients revealed a high frequency of clinically relevant genomic alterations in *CDKN2A*, *FGFR3*, *PIK3CA*, and *ERBB2* in 34%, 21%, 20%, and 17% of patients, respectively (Ross et al., 2016). Another study found that copy number gain in *NOTCH2* was common in BLCA and

that *NOTCH2* promoted growth and invasion of cancer cells by regulating the cell cycle, stemness of cancer, and epithelial-mesenchymal transition (Hayashi et al., 2016). Moreover, the location and expression of lncRNAs are correlated with CNV. A large-scale genomic analysis of cancers found that 995 lncRNAs were located in a focal SCNA region and identified *FAL1* as an oncogenic lncRNA with SCNA that accelerates the growth of tumor cells (Hu XW et al., 2014). Several other studies have reported that protein-coding sequences only occupy less than 2% of the human genome, while over 30% of the genome is affected by SCNAs, indicating that many focal SCNAs in cancer have been mapped to “protein-coding gene desert” regions (Beroukhi et al., 2010; Du et al., 2013; Zack et al., 2013). Collectively, these investigations have shown that, in combination with the diverse mechanisms of lncRNAs, lncRNAs with CNV are closely associated with development of cancers.

In this study, we have identified five novel prognosis-related lncRNAs, *LOC101928489*, *NR2F1-AS1*, *TMEM147-AS1*, *THUMPD3-AS1*, and *LINC01138*, with copy number alterations in BLCA by analyzing datasets in TCGA. Four of these five lncRNAs, *NR2F1-AS1*, *TMEM147-AS1*, *THUMPD3-AS1*, and *LINC01138* were validated in the GEO dataset. *NR2F1-AS1* is an evolutionarily conserved lncRNA that is involved in neurodevelopmental disorders (Ang et al., 2019) and tumor biology in humans. *NR2F1-AS1* is familiarly observed to participate in tumor proliferation, invasion and migration of cancers, including papillary thyroid carcinoma (Guo et al., 2019; Yang et al., 2020), esophageal squamous cell carcinoma (Zhang et al., 2019), and osteosarcoma (Li et al., 2019), through different mechanisms. Huang et al. (2018) also found overexpression of *NR2F1-AS1*, and its targeting on the miR-363-ABCC1 pathway, in oxaliplatin-resistant hepatocellular carcinoma. However, it is far from clear whether copy number changes are associated with *NR2F1-AS1* in tumors. *LINC01138* was found to be a protective factor in our study. However, *LINC01138* has been reported to act as a significant oncogenic driver by interacting with *PRMT5* and enhancing its protein stability to promote tumorigenicity, invasion, and metastasis in hepatocellular carcinoma (HCC) (Li et al., 2018). Similar results were also found in clear cell renal cell carcinoma (Zhang et al., 2018). There are still insufficient data to definitely resolve the

potential role of *LINC01138* in tumorigenesis. Research to date suggests that *THUMPD3-AS1*, by acting as an endogenous sponge of microRNA-543, regulates its target gene *ONECUT2*, and enhances proliferation, relapse, and self-renewal of non-small cell lung cancer (Hu et al., 2019). However, the significance of *THUMPD3-AS1* or *TMEM147-AS1* in BLCA is still unclear. Despite the evidence that lncRNAs seem to be a key factor in prognosis of multiple tumors, the prognostic characteristics of lncRNAs associated with copy number changes remain less clear. These four identified prognostic lncRNAs require further validation from clinical BLCA.

5 Conclusions

In summary, we have identified five CNV-related lncRNAs that are closely correlated with survival prognosis in BLCA patients and we have constructed a valuable risk-score model using these lncRNAs. Our analysis found that all five lncRNAs are involved in regulation of key cancer-related biological functions.

Acknowledgments

This study was supported by the Natural Science Foundation of Guangdong Province (No. 2017A030313898) and the Guangzhou Science and Technology Plan Projects (No. 201707010113), China.

Author contributions

Wenwen ZHONG, Dejuan WANG, and Jianguang QIU contributed to the study design. Bing YAO, Xiaoxia CHEN, Lei YE, Hu QU, Bo MA, and Zhongyang WANG contributed to data analysis. Wenwen ZHONG contributed to writing the paper, Dejuan WANG and Jianguang QIU performed the editing and revision of the manuscript, and provided intellectual support. All authors have read and approved the final manuscript and, therefore, have full access to all the data in the study and take responsibility for the integrity and security of the data.

Compliance with ethics guidelines

Wenwen ZHONG, Dejuan WANG, Bing YAO, Xiaoxia CHEN, Zhongyang WANG, Hu QU, Bo MA, Lei YE, and Jianguang QIU declare that they have no conflict of interest.

This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Ang CE, Ma Q, Wapinski OL, et al., 2019. The novel lncRNA *lnc-NR2F1* is pro-neurogenic and mutated in

- human neurodevelopmental disorders. *Elife*, 8:e41770.
<https://doi.org/10.7554/eLife.41770>
- Beroukhi R, Mermel CH, Porter D, et al., 2010. The landscape of somatic copy-number alteration across human cancers. *Nature*, 463(7283):899-905.
<https://doi.org/10.1038/nature08822>
- Blanche P, Dartigues JF, Jacqmin-Gadda H, 2013. Estimating and comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. *Stat Med*, 32(30):5381-5397.
<https://doi.org/10.1002/sim.5958>
- Bray F, Ferlay J, Soerjomataram I, et al., 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*, 68(6):394-424.
<https://doi.org/10.3322/caac.21492>
- Chappell G, Silva GO, Uehara T, et al., 2016. Characterization of copy number alterations in a mouse model of fibrosis-associated hepatocellular carcinoma reveals concordance with human disease. *Cancer Med*, 5(3):574-585.
<https://doi.org/10.1002/cam4.606>
- Chen G, Wang ZY, Wang DQ, et al., 2013. LncRNADisease: a database for long-non-coding RNA-associated diseases. *Nucleic Acids Res*, 41(D1):D983-D986.
<https://doi.org/10.1093/nar/gks1099>
- Chen XJ, Chang CW, Spoerke JM, et al., 2019. Low-pass whole-genome sequencing of circulating cell-free DNA demonstrates dynamic changes in genomic copy number in a squamous lung cancer clinical cohort. *Clin Cancer Res*, 25(7):2254-2263.
<https://doi.org/10.1158/1078-0432.CCR-18-1593>
- Chen YA, Lemire M, Choufani S, et al., 2013. Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics*, 8(2):203-209.
<https://doi.org/10.4161/epi.23470>
- Cheng LJ, Pandya PH, Liu EZ, et al., 2019. Integration of genomic copy number variations and chemotherapy-response biomarkers in pediatric sarcoma. *BMC Med Genomics*, 12(S1):23.
<https://doi.org/10.1186/s12920-018-0456-5>
- Cibulskis K, Lawrence MS, Carter SL, et al., 2013. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat Biotechnol*, 31(3):213-219.
<https://doi.org/10.1038/nbt.2514>
- Du Z, Fei T, Verhaak RGW, et al., 2013. Integrative genomic analyses reveal clinically relevant long noncoding RNAs in human cancer. *Nat Struct Mol Biol*, 20(7):908-913.
<https://doi.org/10.1038/nsmb.2591>
- Eggers S, DeBoer KD, van den Bergen J, et al., 2015. Copy number variation associated with meiotic arrest in idiopathic male infertility. *Fertil Steril*, 103(1):214-219.
<https://doi.org/10.1016/j.fertnstert.2014.09.030>
- Erriquez J, Becco P, Olivero M, et al., 2015. TOP2A gene copy gain predicts response of epithelial ovarian cancers to pegylated liposomal doxorubicin: TOP2A as marker of response to PLD in ovarian cancer. *Gynecol Oncol*, 138(3):627-633.
<https://doi.org/10.1016/j.ygyno.2015.06.025>
- Feuk L, Carson AR, Scherer SW, 2006. Structural variation in the human genome. *Nat Rev Genet*, 7(2):85-97.
<https://doi.org/10.1038/nrg1767>
- Gudenas BL, Wang J, Kuang SZ, et al., 2019. Genomic data mining for functional annotation of human long noncoding RNAs. *J Zhejiang Univ-Sci B (Biomed & Biotechnol)*, 20(6):476-487.
<https://doi.org/10.1631/jzus.B1900162>
- Guo F, Fu QF, Wang Y, et al., 2019. Long non-coding RNA NR2F1-AS1 promoted proliferation and migration yet suppressed apoptosis of thyroid cancer cells through regulating miRNA-338-3p/CCND1 axis. *J Cell Mol Med*, 23(9):5907-5919.
<https://doi.org/10.1111/jcmm.14386>
- Hänzelmann S, Castelo R, Guinney J, 2013. GSEA: gene set variation analysis for microarray and RNA-Seq data. *BMC Bioinformatics*, 14:7.
<https://doi.org/10.1186/1471-2105-14-7>
- Hayashi T, Gust KM, Wyatt AW, et al., 2016. Not all NOTCH is created equal: the oncogenic role of NOTCH2 in bladder cancer and its implications for targeted therapy. *Clin Cancer Res*, 22(12):2981-2992.
<https://doi.org/10.1158/1078-0432.CCR-15-2360>
- Henrichsen CN, Chaignat E, Reymond A, 2009. Copy number variants, diseases and gene expression. *Hum Mol Genet*, 18(R1):R1-R8.
<https://doi.org/10.1093/hmg/ddp011>
- Hu J, Chen Y, Li X, et al., 2019. THUMP3-AS1 is correlated with non-small cell lung cancer and regulates self-renewal through miR-543 and ONECUT2. *Oncotargets Ther*, 12: 9849-9860.
<https://doi.org/10.2147/OTT.S227995>
- Hu LW, Wu YY, Tan DL, et al., 2015. Up-regulation of long noncoding RNA MALAT1 contributes to proliferation and metastasis in esophageal squamous cell carcinoma. *J Exp Clin Cancer Res*, 34(1):7.
<https://doi.org/10.1186/s13046-015-0123-z>
- Hu XW, Feng Y, Zhang DM, et al., 2014. A functional genomic approach identifies *FAL1* as an oncogenic long noncoding RNA that associates with BMI1 and represses p21 expression in cancer. *Cancer Cell*, 26(3):344-357.
<https://doi.org/10.1016/j.ccr.2014.07.009>
- Hu Y, Wang JL, Qian J, et al., 2014. Long noncoding RNA GAPLINC regulates CD44-dependent cell invasiveness and associates with poor prognosis of gastric cancer. *Cancer Res*, 74(23):6890-6902.
<https://doi.org/10.1158/0008-5472.CAN-14-0686>
- Huang H, Chen J, Ding CM, et al., 2018. LncRNA NR2F1-AS1 regulates hepatocellular carcinoma oxaliplatin resistance by targeting ABCC1 via miR-363. *J Cell Mol Med*, 22(6):3238-3245.
<https://doi.org/10.1111/jcmm.13605>
- Hussain SA, James ND, 2003. The systemic treatment of advanced and metastatic bladder cancer. *Lancet Oncol*, 4(8):489-497.
[https://doi.org/10.1016/S1470-2045\(03\)01168-9](https://doi.org/10.1016/S1470-2045(03)01168-9)
- Kim JS, Chae Y, Ha YS, et al., 2012. Ras association domain family 1A: a promising prognostic marker in recurrent

- nonmuscle invasive bladder cancer. *Clin Genitourin Cancer*, 10(2):114-120.
<https://doi.org/10.1016/j.clgc.2011.12.003>
- Langfelder P, Horvath S, 2008. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, 9:559.
<https://doi.org/10.1186/1471-2105-9-559>
- Ledet EM, Hu XF, Sartor O, et al., 2013. Characterization of germline copy number variation in high-risk African American families with prostate cancer. *Prostate*, 73(6): 614-623.
<https://doi.org/10.1002/pros.22602>
- Li SL, Zheng K, Pei Y, et al., 2019. Long noncoding RNA *NR2F1-AS1* enhances the malignant properties of osteosarcoma by increasing forkhead box A1 expression via sponging of microRNA-483-3p. *Aging*, 11(23):11609-11623.
<https://doi.org/10.18632/aging.102563>
- Li Z, Zhang JW, Liu XY, et al., 2018. The LINC01138 drives malignancies via activating arginine methyltransferase 5 in hepatocellular carcinoma. *Nat Commun*, 9:1572.
<https://doi.org/10.1038/s41467-018-04006-0>
- Liu D, Xu XY, Wen JM, et al., 2018. Integrated genome-wide analysis of gene expression and DNA copy number variations highlights stem cell-related pathways in small cell esophageal carcinoma. *Stem Cells Int*, 2018:3481783.
<https://doi.org/10.1155/2018/3481783>
- Malhotra D, Sebat J, 2012. CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell*, 148(6): 1223-1241.
<https://doi.org/10.1016/j.cell.2012.02.039>
- Martin-Doyle W, Leow JJ, Orsola A, et al., 2015. Improving selection criteria for early cystectomy in high-grade T1 bladder cancer: a meta-analysis of 15215 patients. *J Clin Oncol*, 33(6):643-650.
<https://doi.org/10.1200/JCO.2014.57.6967>
- Martínez-Fernández M, Feber A, Dueñas M, et al., 2015. Analysis of the polycomb-related lncRNAs *HOTAIR* and *ANRIL* in bladder cancer. *Clin Epigenetics*, 7:109.
<https://doi.org/10.1186/s13148-015-0141-x>
- McCarroll SA, Huett A, Kuballa P, et al., 2008. Deletion polymorphism upstream of *IRGM* associated with altered *IRGM* expression and Crohn's disease. *Nat Genet*, 40(9): 1107-1112.
<https://doi.org/10.1038/ng.215>
- Meng QT, Wang KL, Brunetti T, et al., 2018. The *DGCR5* long noncoding RNA may regulate expression of several schizophrenia-related genes. *Sci Transl Med*, 10(472): eaat6912.
<https://doi.org/10.1126/scitranslmed.aat6912>
- Mermel CH, Schumacher SE, Hill B, et al., 2011. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol*, 12(4):R41.
<https://doi.org/10.1186/gb-2011-12-4-r41>
- Mitra AP, Cote RJ, 2009. Molecular pathogenesis and diagnostics of bladder cancer. *Annu Rev Pathol Mech Dis*, 4:251-285.
<https://doi.org/10.1146/annurev.pathol.4.110807.092230>
- Moran VA, Perera RJ, Khalil AM, 2012. Emerging functional and mechanistic paradigms of mammalian long non-coding RNAs. *Nucleic Acids Res*, 40(14):6391-6400.
<https://doi.org/10.1093/nar/gks296>
- Nakamura Y, 2009. DNA variations in human and medical genetics: 25 years of my experience. *J Hum Genet*, 54(1): 1-8.
<https://doi.org/10.1038/jhg.2008.6>
- Ning SW, Zhang JZ, Wang P, et al., 2016. Lnc2Cancer: a manually curated database of experimentally supported lncRNAs associated with various human cancers. *Nucleic Acids Res*, 44(D1):D980-D985.
<https://doi.org/10.1093/nar/gkv1094>
- Nørskov MS, Frikke-Schmidt R, Bojesen SE, et al., 2011. Copy number variation in *glutathione-S-transferase T1* and *M1* predicts incidence and 5-year survival from prostate and bladder cancer, and incidence of corpus uteri cancer in the general population. *Pharmacogenom J*, 11(4):292-299.
<https://doi.org/10.1038/tpj.2010.38>
- Peter S, Borkowska E, Drayton RM, et al., 2014. Identification of differentially expressed long noncoding RNAs in bladder cancer. *Clin Cancer Res*, 20(20):5311-5321.
<https://doi.org/10.1158/1078-0432.CCR-14-0706>
- Potocki L, Bi WM, Treadwell-Deering D, et al., 2007. Characterization of Potocki-Lupski Syndrome (dup(17)(p11.2p11.2)) and delineation of a dosage-sensitive critical interval that can convey an autism phenotype. *Am J Hum Genet*, 80(4):633-649.
<https://doi.org/10.1086/512864>
- Redon R, Ishikawa S, Fitch KR, et al., 2006. Global variation in copy number in the human genome. *Nature*, 444(7118): 444-454.
<https://doi.org/10.1038/nature05329>
- Robinson MD, McCarthy DJ, Smyth GK, 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1): 139-140.
<https://doi.org/10.1093/bioinformatics/btp616>
- Ross JS, Wang K, Khaira D, et al., 2016. Comprehensive genomic profiling of 295 cases of clinically advanced urothelial carcinoma of the urinary bladder reveals a high frequency of clinically relevant genomic alterations. *Cancer*, 122(5):702-711.
<https://doi.org/10.1002/cncr.29826>
- Rubinstein JC, Brown TC, Goh G, et al., 2016. Chromosome 19 amplification correlates with advanced disease in adrenocortical carcinoma. *Surgery*, 159(1):296-301.
<https://doi.org/10.1016/j.surg.2015.09.001>
- Shen RL, Olshen AB, Ladanyi M, 2009. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics*, 25(22):2906-2912.
<https://doi.org/10.1093/bioinformatics/btp543>
- Subramanian A, Tamayo P, Mootha VK, et al., 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci USA*, 102(43):15545-15550.
<https://doi.org/10.1073/pnas.0506580102>

- The Wellcome Trust Case Control Consortium, 2010. Genome-wide association study of CNVs in 16 000 cases of eight common diseases and 3000 shared controls. *Nature*, 464(7289):713-720.
<https://doi.org/10.1038/nature08979>
- Xu HT, Zhu X, Xu ZL, et al., 2015. Non-invasive analysis of genomic copy number variation in patients with hepatocellular carcinoma by next generation DNA sequencing. *J Cancer*, 6(3):247-253.
<https://doi.org/10.7150/jca.10747>
- Yang CJ, Liu Z, Chang XY, et al., 2020. NR2F1-AS1 regulated miR-423-5p/SOX12 to promote proliferation and invasion of papillary thyroid carcinoma. *J Cell Biochem*, 121(2):2009-2018.
<https://doi.org/10.1002/jcb.29435>
- Yu GC, Wang LG, Han YY, et al., 2012. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*, 16(5):284-287.
<https://doi.org/10.1089/omi.2011.0118>
- Zack TI, Schumacher SE, Carter SL, et al., 2013. Pan-cancer patterns of somatic copy number alteration. *Nat Genet*, 45(10):1134-1140.
<https://doi.org/10.1038/ng.2760>
- Zhang X, Wu J, Wu CC, et al., 2018. The LINC01138 interacts with PRMT5 to promote SREBP1-mediated lipid desaturation and cell growth in clear cell renal cell carcinoma. *Biochem Biophys Res Commun*, 507(1-4):337-342.
<https://doi.org/10.1016/j.bbrc.2018.11.036>
- Zhang YW, Zheng AP, Xu RP, et al., 2019. NR2F1-induced NR2F1-AS1 promotes esophageal squamous cell carcinoma progression via activating Hedgehog signaling pathway. *Biochem Biophys Res Commun*, 519(3):497-504.
<https://doi.org/10.1016/j.bbrc.2019.09.015>

Supplementary information

Figs. S1 and S2